

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Facultad de Ingeniería

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Aprendizaje Automático

1er cuatrimestre de 2021

Grupo No. 2 - Move 37

Diego Sodor, Ezequiel Rodriguez, Lucas Dáttoli



Maestría en Explotación de Datos
y Descubrimiento del Conocimiento

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería



1. Resumen

Con el objetivo de realizar un primer acercamiento al análisis de sentimientos y de fijar conocimientos respecto a modelos basados en ensambles de árboles, en el presente trabajo se presentan los resultados de entrenar modelos de Random Forest y AdaBoost para la predicción de sentimientos. Dichos métodos se evalúan sobre conjuntos de audio hablados y cantados del Audio-Visual Database of Emotional Speech and Song (RAVDESS). También se exploró el uso de un Perceptrón Multicapa dada las recomendaciones bibliográficas. Para los 3 modelos entrenados se encontraron resultados satisfactorios, pero Random Forest se destaca por su eficacia en la predicción (unweighted average recall). Además se analizó la robustez del mejor modelo obtenido ante la presencia de ruido en los audios y se notó que la capacidad de clasificación se ve afectada a medida que el ruido aumenta; por otro lado se evaluó el desempeño al estandarizar los datos y se concluyó que esto debería mejorar el rendimiento.

2. Introducción (objetivo y organización del documento)

La clasificación automática del discurso y la música es un tema crítico que surge con la aparición de innovaciones en la tecnología del habla en los dispositivos de nuestra vida cotidiana, como los teléfonos inteligentes y los relojes inteligentes. Mientras que para el reconocimiento automático de discurso existen innovaciones comerciales con gran exactitud enfocadas a usos comerciales, la clasificación de música e información paralingüística más allá del contenido textual, como el estado de ánimo, la emoción, la voz calidad, o personalidad, es un campo muy joven.

El objetivo del presente trabajo es el estudio y análisis de las particularidades de la utilización de algoritmos de ensambles aplicados en casos casi reales de reconocimiento de emociones en señales de sonido. El mismo pretende fijar conceptos de métodos de ensamble; Random Forests; boosting; sobreajuste; tolerancia al ruido. Para esto se utiliza una base de datos que contiene archivos de audio con dos frases expresadas por actores con distintas emociones y niveles de intensidad. Los atributos empleados para entrenar los ensambles son extraídos mediante una librería especializada en la interpretación de voz y música de código abierto. Se explorarán distintas técnicas de división de datos para efectuar el entrenamiento de los algoritmos propuestos. A su vez se evaluarán el rendimiento de cada algoritmo y se ajustarán sus respectivos hiperparámetros para mejorar el desempeño de los mismos en la tarea de reconocimiento de emociones. En una última instancia se estudia la robustez del mejor modelo entrenado ante distintos niveles de ruido y también se analiza el efecto de estandarizar los atributos extraídos de los archivos de audio para entrenar el algoritmo óptimo.

3. Datos

Los datos utilizados son los correspondientes a los archivos de audio del conjunto Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS). Esta base está compuesta por grabaciones de 24 actores profesionales -12 femeninos y 12 masculinos- que pronuncian en inglés y con acento neutral norteamericano, dos declaraciones o discursos léxicamente coincidentes. El discurso incluye expresiones de calma, alegría, tristeza, enojo, temor, sorpresa y disgusto, mientras que la canción contiene expresiones de calma, alegría, enojo y temor. Cada expresión ha sido producida en dos niveles de intensidad emocional -normal y fuerte- y en forma adicional, una expresión neutral.

Estos archivos fueron posteriormente procesados utilizando la librería OpenSMILE, para extraer los atributos del conjunto eGeMAPS que se utilizaron en el entrenamiento de los modelos. OpenSMILE ha sido desarrollada con el objetivo de transformar las señales de audio en atributos que permitan realizar análisis de voz. El conjunto de atributos eGeMAPS ha probado ser útil para diferenciar emociones. Este conjunto¹ contiene 88 atributos relacionados con la frecuencia, la energía/amplitud, y el espectro (Eyben et. al. 2016)² que incluyen parámetros espectrales y dinámicos cuyo éxito para el modelado de emociones ha sido ampliamente probado (Schuller et. al. 2017).

A diferencia de otros conjuntos de atributos -llamados ‘de fuerza bruta’- empleados con anterioridad, los atributos de eGeMAPS fueron seleccionados en función del potencial para indexar los cambios fisiológicos que se producen en la voz asociados a cambios emocionales, la frecuencia y el éxito con el que han sido utilizados previamente y, por último, a su importancia teórica (Eyben et. al. 2016). Adicionalmente, se buscaron atributos acústicos que fueran degradados lo menos posible por el ruido ambiente.

El hecho de que sean los atributos extraídos de los audios los que contienen gran parte de la información asociada a las emociones implica la existencia de un límite importante a la hora de poder entrenar modelos predictivos. Esto se debe a que, a las limitaciones de los modelos tenemos que sumarles las limitaciones de los atributos característicos. **Tampoco resulta menor el hecho de que existan diferencias entre los seres humanos al momento de clasificar la emoción de otro sujeto (Scherer 2003), lo cual implica que las emociones identificadas y clasificadas en el conjunto de audios sean subjetivas y por lo tanto puedan diferir según quién las evalúe.**

4. Metodología

La metodología que se suele emplear para identificar emociones a partir de registros de audio resulta de la aplicación de una combinación de métodos de extracción de características con el entrenamiento de algoritmos para clasificar las emociones a partir de esas características. Resultan relevantes cinco dimensiones desde el punto de vista de la

¹ Eyben et. al. 2016 desarrollaron también un conjunto minimalista de parámetros que contiene sólo 18 descriptores de bajo nivel.

² Una explicación detallada de los atributos seleccionados se puede encontrar en el capítulo 2 de Eyben, 2016.

metodología: la selección de los atributos de audio a analizar, la forma en la que se va a dividir el conjunto de datos para entrenamiento y prueba, los modelos de clasificación a emplear, la forma en la que se van a buscar los hiperparámetros y los mecanismos de normalización de las variables.

Atributos de audio: si bien existen diferentes métodos para extraer los conjuntos de características, tal como señalamos en el apartado anterior, en este caso utilizamos la librería eGeMAPS que fue desarrollada específicamente para extraer atributos asociados a las emociones: *F0*, *volumen*, *proporciones de energía espectral*, *MFCC 1-4*, *paso lineal*, *formante 2-3 de ancho de banda* y *flujo espectral* (Eyben 2016, pág. 132).

División del conjunto de datos: siguiendo las indicaciones recibidas, se eligió Random Forest como modelo de ensamble y se lo entrenó para realizar predicciones dividiendo el dataset de dos maneras: 12-fold cross validation armando los folds de forma aleatoria y Leave-2-Speakers Out. Si bien se obtuvieron mejores resultados con 12-fold CV, se decidió elegir Leave-2-Speakers Out (L2SO) siguiendo las recomendaciones teóricas (Eyben et. al. 2016, Eyben 2016, Gjoreski et. al. 2014, Haider et. al. 2021). Este enfoque es más confiable que usar los datos de la misma persona para entrenamiento y prueba si el modelo está destinado a ser utilizado en personas desconocidas (no incluidas en el conjunto de datos de entrenamiento).

Modelos de clasificación: en lo que respecta a los algoritmos, se suelen emplear Support Vector Machines -SVM- como así también redes neuronales (Eyben et. al. 2016, Eyben 2016, Haider et. al. 2021, Venkataramanan et. A. 2019). En este trabajo, en cambio, utilizamos Random Forest, AdaBoost y Perceptrón Multicapa.

Hiperparámetros: al momento de realizar el ajuste de hiperparámetros se recurrió a la librería Optuna (Akiba et al. 2019), un marco de optimización de hiperparámetros de código abierto para automatizar la búsqueda de los mismos. Se exploró el uso del optimizador Tree-structured Parzen Estimator (TPE). Este es un algoritmo iterativo que usa el historial de hiperparámetros evaluados para crear un modelo probabilístico, que se usa para sugerir el siguiente conjunto de hiperparámetros a evaluar (Bergstra et al. 2011). El funcionamiento detallado del algoritmo excede el alcance de este trabajo. Se prefirió explorar el uso de TPE por sobre el uso de RandomSearch o Grid Search por cuestiones relacionadas al tiempo de procesamiento.

Normalización de variables: se calculó la media y desvío estándar de los atributos para cada actor, y se normalizaron los audios de cada actor utilizando esas estadísticas. Este proceso ha demostrado ser el mejor método para estandarizar los atributos (Eybin 2016, pág 238). De hecho, tal como discutiremos, las predicciones deberían mejorar en los modelos entrenados con variables normalizadas.

La combinación del uso de librerías con relativamente pocos atributos como eGeMAPS con ensambles de árboles como random forest permite obtener modelos con mayor capacidad explicativa que si se utilizan combinaciones que incluyan librerías como ComParE_2016 y modelos como redes neuronales. Esta combinación permite no sólo poder explicar qué atributos resultan ser los más relevantes para clasificar las emociones, sino también verificar que esos atributos sean los mismos que han sido identificados por los ingenieros, psicólogos y lingüistas que los definieron.

No se realizó procesamiento alguno de los atributos extraídos de los archivos de audios en los primeros estadios del trabajo, excepto al momento de entrenar el Perceptrón Multicapa, punto en el que se normalizaron los datos considerando mínimos y máximos de cada atributo (esto facilita el entrenamiento de la red).

5. Resultados

Inicialmente se descargaron los archivos de audios RAVDESS, se extrajeron los atributos eGeMAPS mediante la librería *opensmile* y se generó un dataframe con éstos tanto para los discursos³ (audios “hablados”) como para las canciones⁴ (audios “cantados”). Al entrenar modelos para el reconocimiento de emociones en archivos de audio se exploraron dos estrategias para la división de datos en conjuntos de entrenamiento, validación y prueba.

Para la estrategia que emplea 12-fold cross validation primero se dividió el total de datos en conjuntos de desarrollo y prueba siguiendo una proporción 80:20. Después se crearon de manera aleatoria 12 particiones distintas, las cuales fueron utilizadas para entrenar algoritmos RandomForest y AdaBoost utilizando en cada instancia como conjunto de validación una de las particiones. Como métrica se eligió *Unweighted Average Recall* (UAR), que es una buena métrica de evaluación si varios tipos de emociones tienen una distribución desbalanceada. Esto puede suceder en el entrenamiento debido a las estrategias de validación.

Al entrenar algoritmos RandomForest⁵ y AdaBoost⁶ con los parámetros predeterminados en sus implementaciones de *sklearn* se obtuvo una media de UAR = 0.67 para RandomForest y una media de UAR = 0.37 para AdaBoost en los conjuntos de validación. Además, en la Tabla 1 y Tabla 2, se muestran los resultados asociados al conjunto de prueba. Es evidente que RandomForest es superior a AdaBoost con mejor rendimiento según las métricas de accuracy, precision, recall y F1. Para RandomForest la emoción que supone un mayor desafío es la tristeza, mientras que AdaBoost muestra peor desempeño al identificar disgusto.

³ https://zenodo.org/record/1188976/files/Audio_Speech_Actors_01-24.zip?download=1

⁴ https://zenodo.org/record/1188976/files/Audio_Song_Actors_01-24.zip?download=1

⁵ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

⁶ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

Tabla 1 - Reporte de clasificación para RandomForest 12CV

	precision	recall	f1-score	support
neutral	0.68	0.68	0.68	34
calm	0.74	0.82	0.77	92
happy	0.61	0.58	0.59	64
sad	0.52	0.47	0.49	70
angry	0.71	0.76	0.73	74
fearful	0.71	0.62	0.66	81
disgust	0.53	0.55	0.54	38
surprise	0.71	0.76	0.73	38
accuracy			0.66	491
macro avg	0.65	0.65	0.65	491
weighted avg	0.66	0.66	0.66	491

Tabla 2 - Reporte de clasificación para AdaBoost 12CV

	precision	recall	f1-score	support
neutral	0.32	0.24	0.27	34
calm	0.52	0.70	0.60	92
happy	0.29	0.33	0.31	64
sad	0.23	0.20	0.22	70
angry	0.35	0.45	0.39	74
fearful	0.33	0.14	0.19	81
disgust	0.20	0.05	0.08	38
surprise	0.41	0.79	0.54	38
accuracy			0.37	491
macro avg	0.33	0.36	0.32	491
weighted avg	0.35	0.37	0.34	491

Luego se volvió a entrenar a los algoritmos siguiendo una estrategia de Leave-2-Speakers Out. Cada actor se incluyó dentro una pareja, contando un hombre y una mujer por pareja. La última pareja (actores 23 y 24) se definió como el conjunto de datos de prueba y el resto de las 11 parejas fue utilizada en un esquema de validación cruzada donde utilizando la función *LeaveOneGroupOut* de *sklearn* se designó una pareja como conjunto de validación, y el resto como conjunto de entrenamiento en cada iteración. Nuevamente utilizando los parámetros predeterminados en sus implementaciones de *sklearn* se obtuvo una media de UAR = 0.57 para RandomForest y una media de UAR = 0.36 para AdaBoost en los conjuntos de validación. Los resultados en el conjunto de prueba se exhiben en la Tabla 3 y Tabla 4.

Otra vez se encuentra que RandomForest es superior a AdaBoost con mejor rendimiento en todas las métricas. Con esta estrategia de división de datos, RandomForest tiene problemas para identificar el disgusto, mientras que AdaBoost parece no ser bueno en identificar disgusto y miedo. Como se mencionó en la sección de Metodología, se eligió continuar con la estrategia L2SO, cuyos resultados son menos sesgados al momento de emplear el modelo en personas que no se encuentran incluidas en el conjunto de entrenamiento.

Tabla 3 - Reporte de clasificación para RandomForest L2SO

	precision	recall	f1-score	support
neutral	0.47	0.56	0.51	16
calm	0.68	0.72	0.70	32
happy	0.41	0.41	0.41	32
sad	0.58	0.59	0.58	32
angry	0.79	0.59	0.68	32
fearful	0.62	0.56	0.59	32
disgust	0.55	0.38	0.44	16
surprise	0.50	0.81	0.62	16
accuracy			0.58	208
macro avg	0.57	0.58	0.57	208
weighted avg	0.59	0.58	0.58	208

Tabla 2 - Reporte de clasificación para AdaBoost L2SO

	precision	recall	f1-score	support
neutral	0.21	0.25	0.23	16
calm	0.35	0.53	0.42	32
happy	0.34	0.31	0.33	32
sad	0.36	0.16	0.22	32
angry	0.40	0.78	0.53	32
fearful	0.33	0.06	0.11	32
disgust	0.40	0.12	0.19	16
surprise	0.44	0.69	0.54	16
accuracy			0.37	208
macro avg	0.36	0.36	0.32	208
weighted avg	0.36	0.37	0.32	208

El siguiente paso consistió en ajustar los hiperparámetros de RandomForest y AdaBoost. Para ello se utilizó el optimizador TPE de la librería *optuna* introducido anteriormente. Se buscó maximizar el valor de UAR, métrica que se encuentra respaldada por la literatura asociada al reconocimiento de emoción en el habla. RandomForest demostró una vez más ser mejor que AdaBoost, y los hiperparámetros finales son:

```
bootstrap: False          max_features: 6          min_samples_split: 3
criterion: 'entropy'     max_depth: 21
n_estimators: 266        min_samples_leaf: 2
```

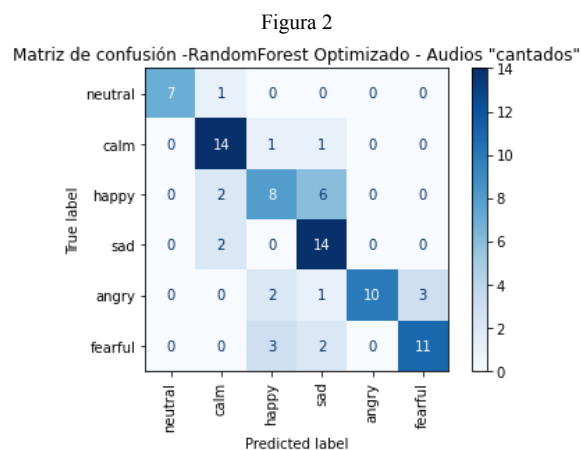
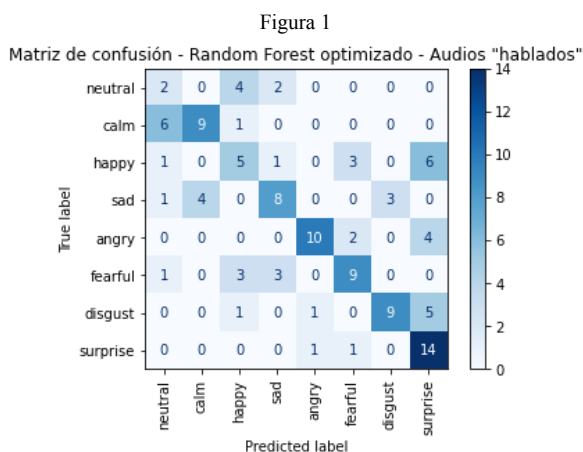
El algoritmo de RandomForest optimizado obtiene una media de UAR = 0.63 en el conjunto de validación, mientras que AdaBoost consigue una media de UAR = 0.39. En el conjunto de prueba RandomForest muestra un accuracy de 0.62, precision de 0.64, recall igual a 0.63 y un F1 de 0.62, superando significativamente a AdaBoost cuyas métricas quedan muy por detrás. Los reportes de clasificación para ambos algoritmos se exponen en la Tabla 5 y Tabla 6. Además se muestran las matrices de confusión de RandomForest para los audios “hablados” y “cantados” en la Figura 1 y Figura 2. Tanto el reporte de clasificación en todo el conjunto de prueba como las matrices de confusión muestran que RandomForest encuentra como mayor desafío identificar alegría.

Tabla 5 - Reporte de clasificación para RandomForest optimizado

	precision	recall	f1-score	support
neutral	0.50	0.56	0.53	16
calm	0.72	0.72	0.72	32
happy	0.46	0.41	0.43	32
sad	0.58	0.69	0.63	32
angry	0.91	0.62	0.74	32
fearful	0.69	0.62	0.66	32
disgust	0.75	0.56	0.64	16
surprise	0.48	0.88	0.62	16
accuracy			0.62	208
macro avg	0.64	0.63	0.62	208
weighted avg	0.65	0.62	0.63	208

Tabla 6 - Reporte de clasificación para AdaBoost optimizado

	precision	recall	f1-score	support
neutral	0.38	0.31	0.34	16
calm	0.38	0.75	0.51	32
happy	0.31	0.31	0.31	32
sad	0.43	0.09	0.15	32
angry	0.52	0.50	0.51	32
fearful	0.50	0.56	0.53	32
disgust	1.00	0.06	0.12	16
surprise	0.44	0.69	0.54	16
accuracy			0.42	208
macro avg	0.50	0.41	0.38	208
weighted avg	0.47	0.42	0.39	208



Para contrastar nuestros resultados con la literatura de reconocimiento de emociones en archivos de sonido, calculamos la importancia de los atributos utilizados en el RandomForest. Encontramos que entre los 10 atributos con mayor importancia, 8 corresponden a variantes de F0, flujo espectral y volumen, que son atributos especialmente diseñados para el reconocimiento de emociones (Eyben 2016, págs. 131-133). Utilizando los 60 mejores atributos para reentrenar el modelo se obtiene un rendimiento similar.

Para familiarizarnos con el uso de redes neuronales, también se entrenó un Perceptrón Multicapa. La red optimizada obtuvo una media de UAR = 0.51 en el conjunto de validación. Como la implementación de este algoritmo era opcional, el reporte de clasificación y la matriz de confusión se exponen en la Tabla A.1 y Figura A.1 del Apéndice A, respectivamente. El RandomForest sigue siendo el modelo que consigue el mejor rendimiento. Sin embargo, es importante aclarar que el espacio de hiperparámetros para la red neuronal es mucho más amplio, y es muy probable que exista una configuración de parámetros no encontrada que supere a RandomForest.

Mediante el análisis de distintos niveles de ruido sintético uniforme introducidos a los audios de prueba, se notó un drástico descenso en la performance de Random Forest con el aumento del ruido (Ver figuras B.1 y B.2 del Apéndice B). Lo que permite concluir que el modelo entrenado no es robusto ante el ruido.

En una última etapa se normalizaron los archivos de audio y entrenamiento por actor y se volvió a entrenar el RandomForest optimizado con estos datos. Se encontró que el rendimiento del modelo mejora levemente para accuracy obteniendo un valor de 0.63, pero empeora para precisión con un valor de 0.63, recall igual a 0.62 y un F1 de 0.61. La variación de las métricas se dio en centésimas. Se esperaba que al estandarizar los audios se eliminaran particularidades de cada actor, y de esta manera el modelo aprendiera los patrones comunes asociados a las emociones y no a la forma que tiene cada actor de expresarlas.

6. Conclusiones

En este primer acercamiento al reconocimiento de emociones en el habla (SER) se investigó el uso de atributos extraídos a partir de señales de audio. Particularmente, se evaluó la efectividad de diferentes algoritmos de aprendizaje supervisados, como Random Forest, AdaBoost y Redes Neuronales utilizando distintos métodos para optimizar su efectividad. Los resultados obtenidos indican que el modelo que mejor performa en términos de la métrica elegida, el recall, es Random Forest. Gracias a esto podemos concluir que este método de clasificación permite obtener resultados satisfactorios.

Durante el desarrollo del trabajo se tomó conciencia de ciertas limitaciones y desafíos del uso de RAVDESS para hacer el análisis. Hay un número limitado de conjuntos de datos de habla con sentimientos etiquetados, lo que perjudica en el entrenamiento y evaluación de los modelos utilizados. Por otro lado, la generalización de los modelos podría optimizarse teniendo una base de datos más grande con mayor variedad de lenguajes, acentos y culturas.

Se aplicaron distintas técnicas para estimar el error de generalización y optimizar hiperparámetros, entre los que se pueden nombrar: 12-Fold Cross Validation y Leave-OneGroup-Out Cross Validation. Y además se probaron distintos enfoques recomendados en bibliografía sobre SER, como la estandarización por actor de los atributos extraídos del audio que usualmente permite obtener mejores resultados. La introducción de distintos niveles de ruido en las señales de audio del conjunto de prueba permitió ver que el modelo entrenado de Random Forest resultó poco robusto ante el mismo. Los descubrimientos obtenidos a través de la investigación realizada permitieron notar que algunas emociones pueden ser más difíciles de distinguir en el audio, como es el caso de "alegría". Estos resultados podrían complementarse con el análisis de textos o de expresiones faciales en caso de encontrarse disponibles.

Con la evidencia obtenida consideramos que tanto los modelos de aprendizaje de ensambles empleados y las redes neuronales resultan efectivos en este ámbito. Finalmente, consideramos que modelos de redes neuronales más avanzadas podrían llegar a tener buenos resultados, debido a esto se alienta la investigación del uso de redes neuronales recurrentes y convolucionales.

7. Referencias Bibliográficas

- Akiba, T., Sano, S., Yanase, T., Ohta, T, and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. ArXiv, abs/1907.10902.
- Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. Advances in Neural Information Processing Systems 24.
- Eyben, F. (2016). Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Switzerland. Springer.
- Eyben, F. et al. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, vol. 7, no. 02, pp. 190-202.
- Gjoreski, M., Gjoreski, H., & Kulakov, A. (2014). Machine Learning Approach for Emotion Recognition in Speech. Informatica (Slovenia), 38.
- Haider, F., Pollak, S., Albert, P., Luz, S. (2021). Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. Computer Speech & Language, Volume 65.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer Texts in Statistics. New York, NY: Springer.
- Schuller, B., Batliner, A., D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1-2)
- Seni, G., Elder J. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan & Claypool.
- Venkataramanan, K., Rajamohan, H. (2019). Emotion Recognition from Speech. ArXiv, abs/1912.10458.

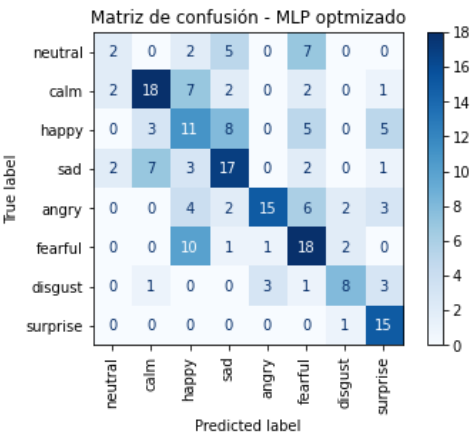
Apéndice A

Resultados de Perceptrón Multicapa

Tabla A.1: Reporte de clasificación para MLP optimizado

	precision	recall	f1-score	support
neutral	0.33	0.12	0.18	16
calm	0.62	0.56	0.59	32
happy	0.30	0.34	0.32	32
sad	0.49	0.53	0.51	32
angry	0.79	0.47	0.59	32
fearful	0.44	0.56	0.49	32
disgust	0.62	0.50	0.55	16
surprise	0.54	0.94	0.68	16
accuracy			0.50	208
macro avg	0.51	0.50	0.49	208
weighted avg	0.52	0.50	0.49	208

Figura A.1:



Hiperparámetros

hidden_layer_sizes: (144,111)
activation: 'tanh'

learning_rate: 'constant'
learning_rate_init: 0.0047

solver: 'adam'
max_iter: 289

Apéndice B

Efectos del ruido en el conjunto de prueba

Figura B.1:



Figura B.2:

